Archive API System Overview Manual

From Intranet

Back to Archive Retrieval Server Level 1 design or the Archive Retrieval Server Development Notes

Contents

Introduction

The primary data holdings at the National Geophysical Data Center (NGDC) are too large to all be held on disk drives for rapid access (also known as "online" or "spinning disk"). Some portion of the data are written to tape and are not normally kept as files on disk drives--these are copied back to disk when needed and may be erased after use so that the disk space can be reused for other files. A web service known as the **Archive Retrieval Server** has been developed at NGDC to automate and control access to these data holdings on tape. This software is sometimes refered to as an API (Application Programming Interface) since this is a web service intended to be used by other programs.

In compliance with National Archive and Records Administration (http://www.archives.gov/about/info-qual/guidelines.html) (NARA) guidlines, all data that are officially "archived" are stored to high capacity tape along with at least one backup copy to preserve data integrity and allow off site backup. To handle the large volume of data NGDC uses a robotic tape library known as the Scalar *Model #*, a product of Quantum (http://www.quantum.com/Products/TapeLibraries/Index.aspx) . This hardware system was formerly known as the "ADIC" system before that company was purchased Quantum. The Quantum software being used is StorNext (http://www.quantum.com/ServiceandSupport/SoftwareandDocumentationDownloads/SNMS/Index.aspx) 4.1. This software manages backup, disk to tape migration, cleanup policy and copy duplication, but does not provide job management for data requests.

Note: after Quantum bought the ADIC company, the "ADIC" name has been dropped from their product names. These documents often refer to the "ADIC" name, but these are being phased out to use the term "Archive" instead. Also programs such as adicJobStatus and adicRetrievalServer are being renamed archiveOrderStatus and archiveRetrievalServer. Some pages in the NGDC intranet wiki have ADIC in their titles, and these will be eventually renamed.

Goals

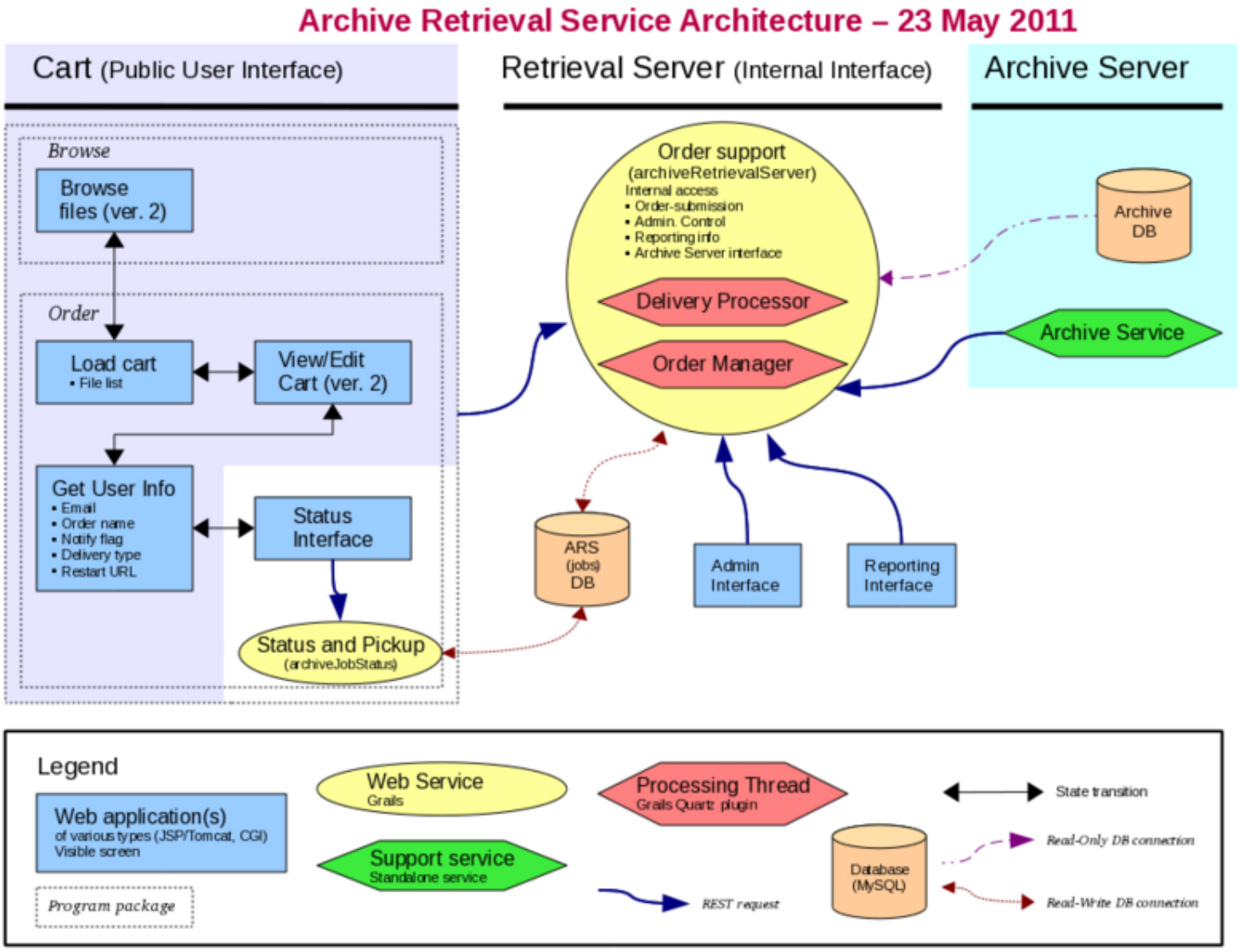
The purpose for the Retrieval Server software developed at NGDC is to provide prioritized data request management. This appears as a single client to the Quantum tape library so that a single point of control is available to manage load on the system. In order to achieve this, the Retrieval Server software has features built on top of the StorNext software to support the following goals:

- Track jobs in support of a shopping cart style front end
  - Provides a web service using REST protocols
- Prioritize jobs, generally processing jobs in the order they were received within a priority group
- Prevent a single user from monopolizing resources, jobs at the same priority by other users are shuffled into the multiple

requests of heavy users, even if received later (this will be implemented later if needed)
- Control access to files under "nonpublic" and "private" directories
- Limit the size of orders according to parameters for size and number of files
- Segment large orders into limited numbers of files per tape retrieval request
- Pause if temporary or final destination disk space becomes too low
- Be able to pre-check files for existence in an inventory database before retrievals are requested from the tape library
- Be able to look for data in archives outside NGDC such as the NEAAT archives in CLASS
- Notify users when a job completes via email
- For external orders:
    - Package retrieved files for a single order as a tar file on a public-facing server for ease of download
    - Provide a status/pick-up web application for users to see their order status, retrieve their completed order, or cancel an order queued for processing.
- For internal orders:
    - Place retrieved files into a chosen nfs mounted delivery area
    - Allow enhanced access to "nonpublic" files
- Provide a Job monitoring and administrative control panel which can
    - View and edit current jobs, including their status, priority, and delivery destination
    - View and edit system parameters
    - View and edit disk delivery destinations
    - Monitor system health, disk and database usage
    - Calculate statistics on usage
    - Pause or shutdown processing

The administrative features are accessed only internally by a web application that requires a login. The prioritizing, file validation and segmenting is automated so that users of the service typically don't need to set any parameters in this regard. Jobs and a number of parameters are stored in a database so that the state of the system is maintained across system downtime.

System Architecture



The white portion of this diagram indicates the features provided by the Retrieval Server to manage data requested data orders as a sequence of jobs. The area shaded purple represents the front end software that makes use of the Retrieval Server, interacting with users directly to present inventory and collect data orders. The aqua area describes the Archive Server software that interacts directly with the ADIC software to perform retrievals.

Specifications and Performance

Once testing is complete with sample data, we will have details on system throughput.

See Also

- Archive API Programmers Manual
- Archive Retrieval Server Level 1 design
- Archive Retrieval Server Development Notes